Design Space Exploration of Distributed ML

Course Project
CSCE 790
(Machine Learning Systems)

Project description

- How the choice of configuration parameters in distributed ML setting affect training time?

Project goal

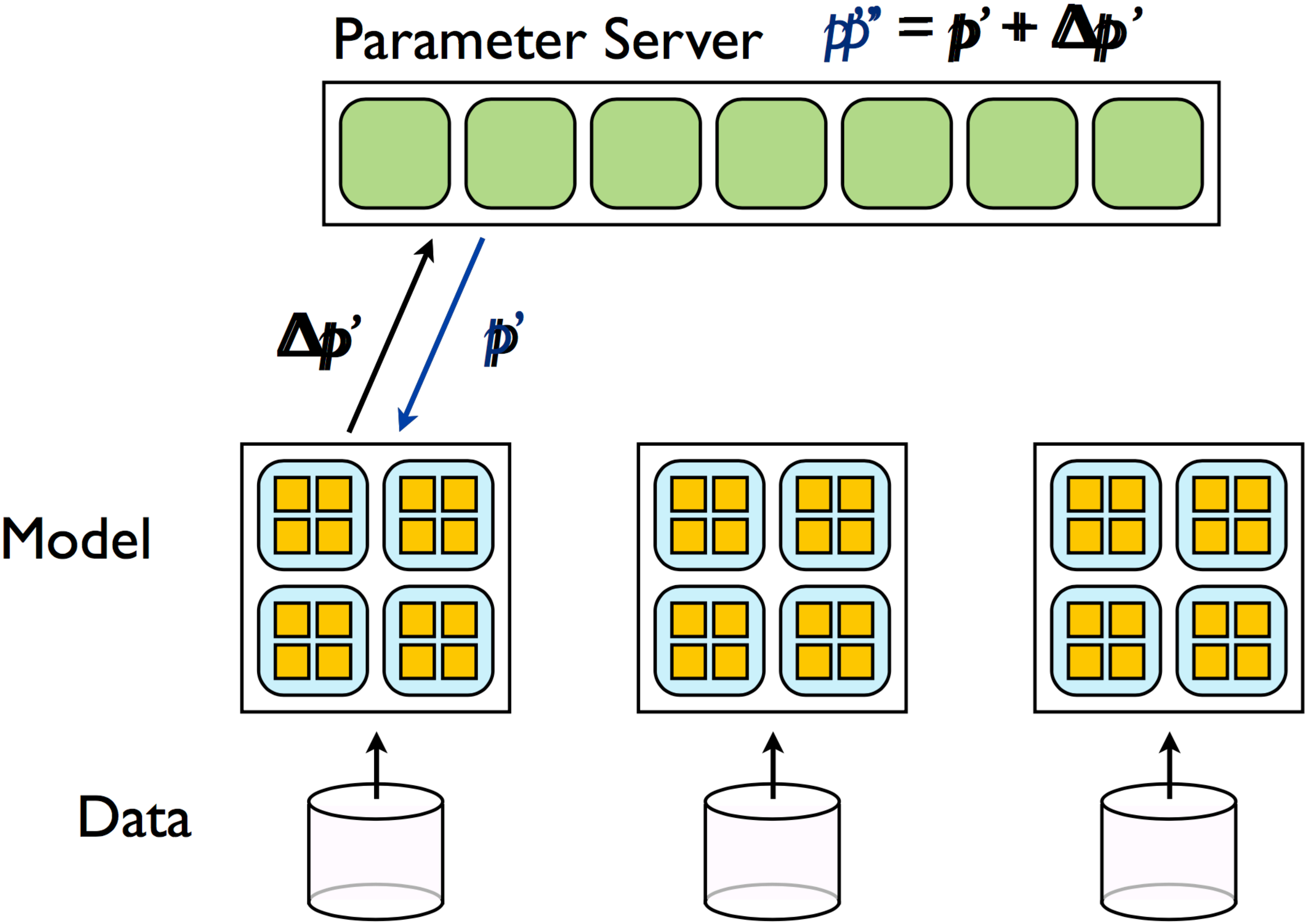
- The aim of the project is to perform design space exploration of distributed ML.

- The goal is to understand how the choice of configurations in the training environment can influence training time of DNNs.

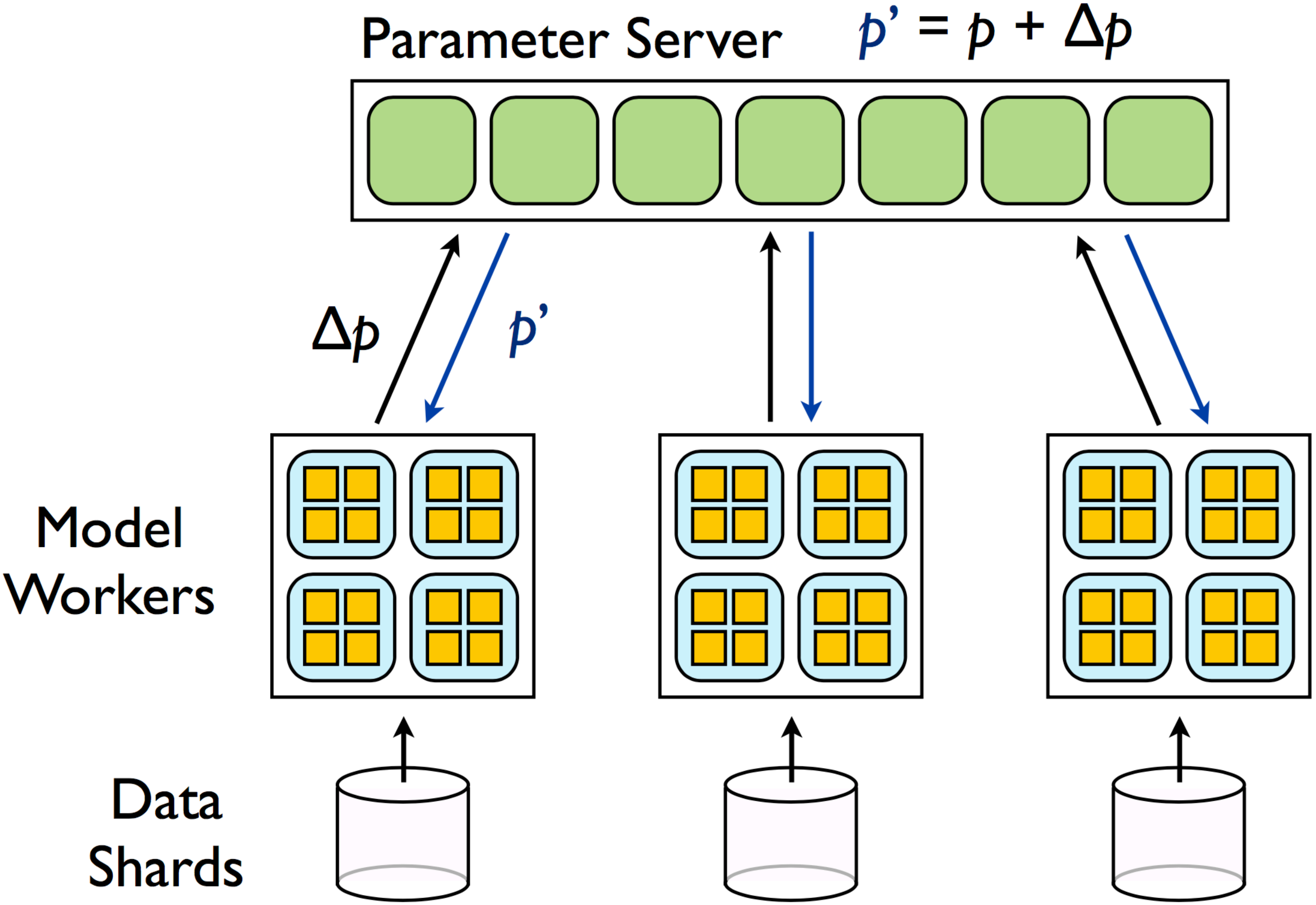
Distributed ML

- There are various approaches to accelerate training of DNNs.
    - Data parallelism, where you shard training data across multiple nodes.
    - Model parallelism, where you split the model across multiple nodes.
    - The focus of this project is on the “data parallelism”.

Data parallelism



Data parallelism



Setting up the experimental environment

- First you need to setup your experimental environment
  - 2 nodes
  - Each with 1-2 GPU
  - I may be able to provide a simple environment, contact me once you created your team.
  - You may use your own GPU servers.

Deciding the configuration space

- You need to select few configuration options that affect performance, e.g., :
  - Number of parameter servers
  - Number of worker nodes
  - Communication protocol
  - Buffer size
  - etc

Selecting specific DNN architectures

- Select few pre-trained DNN architectures that fit onto your hardware platform, e.g.:
  - Any pre-trained CNN architecture
  - Use available implementations, e.g.,: https://github.com/tensorflow/benchmarks

Deciding about workload

- Choose 2 different workloads from existing datasets, e.g. UCI repository, or other available datasets
  - Image
  - Time-series
  - Text
  - etc.

Start measurements

- Once you decided about the configuration space, you need to determine the configurations that you want to measure.
- At this stage you need to discretize the continuous variables.
- And think about using a sampling strategy, e.g., random sampling, or possibly Full factorial design
  - https://en.wikipedia.org/wiki/Design_of_experiments
- You need to measure training time for each configuration

Analyzing data

- Once you measured configurations, you need to dig into data and find interesting trends.
    - You could look into optimal configurations
    - You could find whether the optimal configurations in one DNN architecture is also optimal in other architectures, if not dig into and find out why.
    - You could look into correlation measures across different workloads
    - You may want to have a look at this to get some idea what kinds of analyses you may want to perform: https://arxiv.org/pdf/1804.01138.pdf

Final point

- Use your creativity when it comes to analyzing the results, try to surprise me!

- If you find a very interesting observations and dig into it by providing some insight, you will then get a good score!

- If you also produce very good results, you may also want to think about a potential paper, it’s optional, but I strongly recommend it.